MuDCoD: Multi-Subject Community Detection in Dynamic Gene Networks

Ali O. B. Şapcı[1], Shan Lu[2], Oznur Tastan[1,*], Sündüz Keleş[2,3,*]

1: Faculty of Engineering and Natural Sciences, Sabancı University, Turkey
2: Department of Statistics, University of Wisconsin, Madison, WI, USA
3: Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

Motivation

- Single-cell RNA-seq datasets across multiple individuals and time points are now routinely generated for different conditions [1].

Jerber-2021 Dataset:

- The Jerber-2021 dataset [1] contains single-cell RNA-Seq data of cells from 215 iPSC lines derived from the Human Induced Pluripotent Stem Cell Initiative (HipSci) and differentiating toward a mid-brain neural fate.
- After pre-processing, we considered the scRNA-seq count matrices from day-32 and day-50, from 16, 22, and 8 donors, respectively, for DA, Sert, and Epen1 cell types. Gene-gene networks for each donor at each time point are constructed based on Pearson correlation matrices.

- Analysis of personalized dynamic gene networks constructed from these datasets could unravel subject-specific network-level variation critical for explaining phenotypic differences.

Problem definition

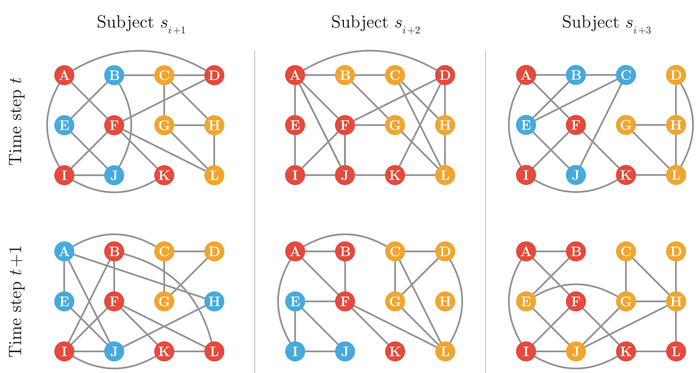


Fig. 1: A schematic for multi-subject dynamic gene networks. There is a gene-gene network for each subject at each time step. The node-set is identical in the networks, and the edges vary among both the subjects and across the time steps. MuDCoD assumes communities change smoothly across both the subject and the time dimensions.

- We define a multi-subject dynamic gene co-expression network for discrete time steps $t = 0, \ldots, T-1$ and for subject $s = 0, \ldots, S-1$ as a time series of undirected and unweighted graphs $\mathcal{G}_{s1}, \ldots \mathcal{G}_{sT}$ for each subject s.
- Given a multi-subject dynamic gene co-expression network, we aim to infer the *communities* for each time point and subject.

Static spectral clustering:

$$L = D^{-1/2} A D^{-1/2} \text{ where } D_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\deg(v_i)$ denotes the degree of node i and A is the adjacency matrix of the $\mathcal{G}$.
Let K be fixed, and $V_{st} \in \mathbb{R}^{G \times K}$ denote a matrix with columns corresponding to the K leading eigenvectors of L_{st}. A baseline strategy for inferring communities separately at each snapshot of time step and for each individual by clustering on V_{st}.

MuDCoD formulation

- PisCES [3] applies smoothing to the eigenvectors of L_{st} across time dimensions.
- MuDCoD applies eigenvector smoothing across both the subject and the time dimensions to promote signal sharing. Let $U_{st} = V_{st} V_{st}^T$ be the projection matrix onto the column space of V_{st} . Define mean projection:

$$\left[\bar{U}_t\right]_{ij} = \frac{1}{S} \sum_{s=0}^{S-1} \left[\overline{U}_{st}\right]_{ij}. \quad (2)$$

where $\overline{U}_{st}$ is the smoothed version of U_{st}. In order to estimate $\overline{U}_{st}$, we propose the following optimization problem;

Smoothness over time and among subjects:

$$\min_{\substack{\overline{U}_{st} \\ s=0,\ldots,S-1 \\ t=0,\ldots,T-1}} \sum_{t=0}^{T-1} \left(\|U_{st} - \overline{U}_{st}\|_F^2 + \beta \|\overline{U}_{st} - \bar{U}_t\|_F^2 \right) + \alpha \sum_{t=0}^{T-2} \|\overline{U}_{st} - \overline{U}_{s(t+1)}\|_F^2 \quad (3)$$

$$\text{subject to } \overline{U}_{st}, \in \left\{ VV^T \ : \ V \in \mathbb{R}^{G \times K}, \ V^T V = I \right\} \ \forall s, \forall t.$$

- $\alpha \|\overline{U}_{st} - \overline{U}_{s(t+1)}\|_F^2$ enforces smoothness over the time dimension.
- $\beta \|\overline{U}_{st} - \bar{U}_t\|_F^2$ constrains subject-specific variation from the mean time-dependent projection matrix $\bar{U}_t$.

We propose to solve this non-convex optimization problem with the following iterative method:

$$\overline{U}_{st}^{\ell+1} = \Pi_K \left(\alpha \overline{U}_{s(t-1)}^{\ell} + U_{st} + \alpha \overline{U}_{s(t+1)}^{\ell} + \beta \bar{U}_t^{\ell} \right), \ t = 1, \ldots, T-2 \quad (4)$$

$$\Pi_K(M) = \sum_{k=1}^{K} v_k v_k^T, \quad (5)$$

where $v_1, \ldots, v_k$ are the K leading eigenvectors of M.

- We allow K to be unknown and possibly vary over time. We utilize the eigengap statistics to select the number of modules, and α and β are chosen with a re-sampling-based cross-validation strategy by [2].

MuDCoD discovers revealing gene modules

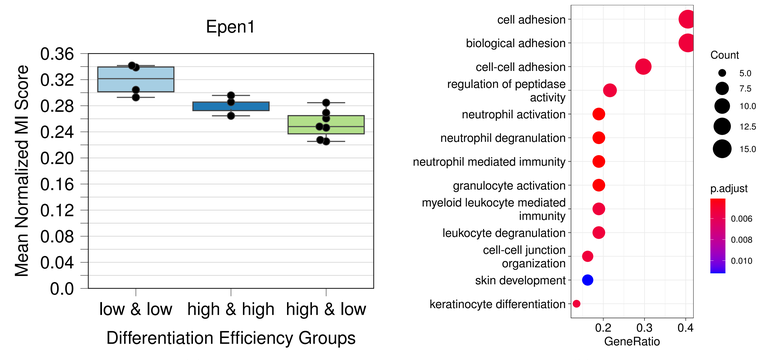
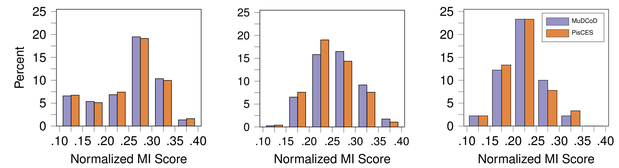


Fig. 2: (a) Mean normalized MI score comparison within and between the low and high differentiation efficiency groups of Epen1 cells. Each point stands for a donor and, the y-axis denotes the mean of normalized MI scores between that donor and other donors in the corresponding group. (b) Set of enriched biological processes for a co-occurrent gene set that is specific to low differentiation efficiency donors' communities in Epen1 cells at day-52. Displayed are significant biological processes (adjusted p-value ≤ 0.05) of one of the largest co-occurrent gene sets (40 genes).

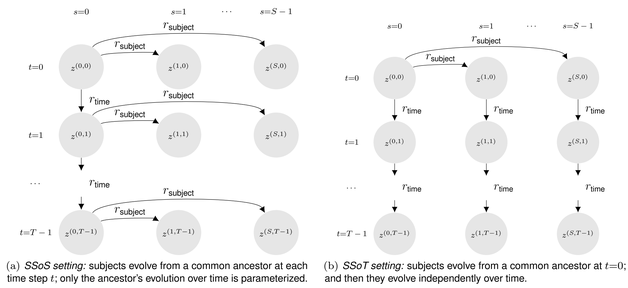
MuDCoD shares signal among subjects



(a) NMI between subjects on day-30. (b) NMI between subjects on day-52. (c) NMI between day-30 and day-52 for each subject.

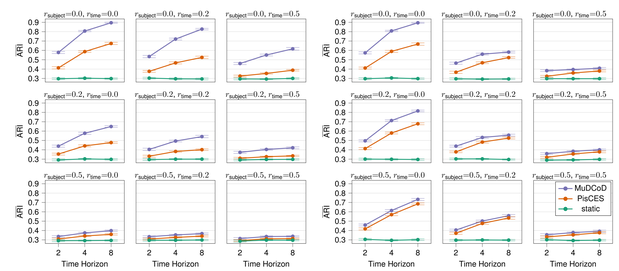
Fig. 3: Normalized MI scores between the inferred modules on day-30 and on day-52 for each subject w.r.t. the differentiation efficiency of the corresponding subject. The similarities between the modules inferred on day-30 and day-52, quantified with the NMI score, tend to decrease with increasing differentiation efficiency.

Performance comparison with simulations



(a) *SSoS setting:* subjects evolve from a common ancestor at each time step t; only the ancestor's evolution over time is parameterized. (b) *SSoT setting:* subjects evolve from a common ancestor at $t=0$; and then they evolve independently over time.

Fig. 4: Multi-subject dynamic degree corrected block models (MuS-Dyn-DCBM) for the two proposed settings.



(a) *The SSoS setting.* (b) *The SSoT setting.*

Fig. 5: Evaluation under two different MuS-Dyn-DCBM settings, (a) SSoS and (b) SSoT. Network size G=500, # of class labels K=10, in-cluster and out-cluster density parameters $p_{\text{in}} = (0.2, 0.4)$ and $p_{\text{out}} = 0.1$, # of subjects S=8, and # of time points $T \in \{2, 4, 8\}$. x-axis denotes T, and y-axis is the mean ARI of the inferred modules for all subjects and time steps.

Conclusion

MuDCoD enables robust inference for identifying time-varying personalized gene modules by leveraging shared signals among the subjects. The implementation is publicly available at GitHub, (scan QR-code).

References

[1] HipSci Consortium et al. "Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation". en. In: *Nature Genetics* 53.3 (Mar. 2021), pp. 304–312.

[2] Tianxi Li, Elizaveta Levina, and Ji Zhu. "Network cross-validation by edge sampling". In: *Biometrika* 107.2 (Apr. 2020), pp. 257–276.

[3] Fuchen Liu et al. "Global spectral clustering in dynamic networks". In: *Proceedings of the National Academy of Sciences* 115.5 (2018), pp. 927–932.